

CURRENT CONCEPTS REVIEW

USER'S GUIDE TO THE ORTHOPAEDIC LITERATURE: HOW TO USE AN ARTICLE ABOUT A SURGICAL THERAPY

BY MOHIT BHANDARI, MD, MSC, GORDON H. GUYATT, MD, MSC, AND MARC F. SWIONTKOWSKI, MD

Investigation performed at the Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Center, Hamilton, Ontario, Canada, and the Department of Orthopaedic Surgery, University of Minnesota, Minneapolis, Minnesota

- We suggest a three-step approach when using an article from the surgical literature to guide your patient care: (1) assess whether the study can provide valid results, (2) review the results, and (3) consider how the results might be applied to your patient.
- Randomization, concealment, intention-to-treat analysis, similarity of patients for known prognostic factors, blinding of patients and outcome assessors, and completeness of follow-up are important guides to study validity.
- The 95% confidence interval around the treatment effect is a measure of precision.
- Consider whether all of the clinically important outcomes were reported and whether the likely benefits of treatment outweigh the potential harm and costs.

Clinical Scenario

You are an orthopaedic surgeon who is called to the emergency department to evaluate and treat a fifty-five-year-old woman with a displaced fracture of the distal aspect of the right radius. She tells you that she fell on her outstretched right hand after slipping on the kitchen floor. Her medications include L-thyroxine and alendronate. On examination, she has an obvious deformity of the wrist and no evidence of neurovascular compromise. Plain radiographs demonstrate a dorsally tilted and comminuted distal radial fracture with no extension into the joint.

You believe that the patient's age and the displacement of the fracture warrant a closed reduction in the operating room. One of your partners who is passing through the emergency department agrees with your assessment and comments that dorsally comminuted fractures tend to be very unstable. Moreover, she suggests that the new "bone cements" on the market might be ideal for preventing secondary instability fol-

lowing a closed reduction. She urges you to find a report of a recent randomized trial, which she recalls having read, in the literature.

Intrigued by your colleague's proposal, you tell her that you will search the literature for articles on calcium-phosphate-based bone-cement materials and will use the information to formulate a plan by the time that your patient is taken to the operating room. The operating-room charge nurse tells you that there are three other cases ahead of yours, which will delay your case by approximately five hours.

The Literature Search

You begin by formulating your question: in patients with displaced distal radial fractures, what is the impact of injectable bone cement on malunion rates compared with that of no treatment? Since the study that you are seeking was published within the last couple of months, you begin with an Internet-based PubMed search, using a so-called "clinical

This article is the first in a series designed to help the orthopaedic surgeon use the published literature in practice. We start with the topic of the controlled trial, which is widely considered to be the study design of greatest utility because of its systematic attempts to limit bias. Controlled trials are difficult to perform in surgery but are being conducted with increasing frequency.

query” and randomized trial sensitivity filter with the following keywords: “fracture” and “calcium phosphate”. This search yields only forty articles, one of which is evidently your target¹ and a second that also seems very relevant².

The first article that you identify is a report of a randomized trial of 110 patients with displaced distal radial fractures who were treated with or without an injectable calcium-phosphate cement (Norian SRS)¹. The second article is a report of a randomized trial of 249 long-bone fractures that were treated with internal or external fixation supplemented either with a collagen-calcium phosphate material or with autogenous bone graft².

The Guide

Most surgical interventions have inherent benefits and associated risks. Before implementing a new therapy, you should ascertain its benefits and risks and assure yourself that the resources consumed in the intervention will not be exorbitant. We suggest that you employ a three-step approach when using an article from the surgical literature to guide your patient care: (1) assess whether the study can provide valid results (internal validity), (2) review the results, and (3) consider how the results might be applied to your patient (generalizability) (Table I).

Validity

Did experimental and control groups begin the study with a similar prognosis?

Were patients randomized?

During the 1970s and early 1980s, surgeons frequently performed extracranial-intracranial bypass (anastomosis of a branch of the external carotid artery, the superficial temporal, to a branch of the internal carotid artery, the middle cerebral). They believed that this prevented strokes in patients who had symptomatic cerebrovascular lesions that were otherwise surgically inaccessible. Studies comparing outcomes among non-randomized cohorts of patients who, for various reasons, did or did not undergo this operation fueled this conviction. These studies suggested that patients who underwent surgery fared much better than those who did not. However, to the investigators' surprise, a large multicenter trial in which patients were allocated to surgical or medical treatment with use of a process analogous to flipping a coin (a *randomized controlled trial*) demonstrated that the only effect of surgery was to increase adverse outcomes in the immediate postsurgical period³.

Randomized trials have led to other surprising findings that have contradicted the results of less rigorous trials. For example, one randomized trial demonstrated that steroid injections do not ameliorate facet-joint back pain⁴, and several others showed that a variety of initially promising drugs increased mortality in patients with heart disease⁵⁻⁹. Such surprises frequently occur when treatments are assigned by random allocation rather than by the conscious decisions of clinicians and patients.

Investigators who study orthopaedic treatments attempt to determine the impact of an intervention on events such as

TABLE I User's Guide to Randomized Trials in Orthopaedics

Validity

Did experimental and control groups begin the study with a similar prognosis?

Were patients randomized?

Was randomization concealed?

Were all patients analyzed in the groups to which they were randomized?

Were patients in the treatment and control groups similar with respect to known prognostic factors?

Did experimental and control groups retain a similar prognosis after the study started?

Blinding

Did investigators avoid effects of patient awareness of allocation: were patients blinded?

Were aspects of care that affect prognosis similar in the two groups: were clinicians blinded?

Was outcome assessed in a uniform way in experimental and control groups: were those assessing outcome blinded?

Was follow-up complete?

Results

How large was the treatment effect?

How precise was the estimate of the treatment effect?

Applicability

Can the results be applied to my patient?

Were all clinically important outcomes considered?

Are the likely treatment benefits worth the potential harm and costs?

nonunion, infection, and death; these occurrences are referred to as the trial's *target outcomes* or *target events*. The patient's age, the underlying severity of the fracture, the presence of comorbid conditions, health habits, and a host of other factors (*prognostic factors* or *determinants of outcome*) typically determine the frequency with which a trial's target outcome occurs. If prognostic factors—either those that we know about or those that we do not—prove to be unbalanced between a trial's treatment and control groups, the outcome will be biased, resulting in either an underestimation or an overestimation of the treatment effect. Since known prognostic factors often influence clinicians' recommendations and patients' decisions about treatment, observational studies often yield misleading results. Typically, observational studies tend to show larger treatment effects than do randomized trials¹⁰⁻¹³, although systematic underestimation of treatment effects may also occur¹⁴.

The disadvantage of randomization in surgical trials is that individual surgeons may not have equal experience or skill in performing the two treatments to be studied. This presents an ethical dilemma when two beneficial treatment op-

tions for a musculoskeletal condition are available, as patients assigned to different treatment arms may not have the same opportunity to receive the best care. In addition, high-quality randomized trials are expensive to conduct, and the results are often not available for several years until follow-up is complete. However, it is not only the trial's design that needs to be satisfactory but also the actual conduct of the trial as it affects each individual patient. Ultimately, it is up to the clinical investigators to ensure that patients do not suffer as a result of the clinical research. The power of randomization is that treatment and control groups are far more likely to be balanced with respect to both the known and the unknown determinants of outcome.

Randomization does not always achieve the investigators' goal of having groups with a similar prognosis. Investigators may make mistakes that compromise randomization. For example, randomization will be compromised if those who determine eligibility are aware of the treatment arm to which the patient will be allocated or if patients' results are not analyzed in the group to which they were allocated.

Was randomization concealed?

In 1996, a group of Australian investigators reported a randomized trial of open compared with laparoscopic appendectomy¹⁵. The trial ran smoothly during the day. At night, however, the attending surgeon's presence was required for the laparoscopic but not the open procedure, and the limited operating-room availability made the longer, laparoscopic procedure an annoyance. Reluctant to call in the consultant and, particularly, specific senior colleagues, residents sometimes adopted a practical solution. When an eligible patient appeared, the resident checked the attending staff and the operating-room line-up and, depending on the situation, held the translucent envelopes up to the light. As soon as an envelope that dictated an open procedure was identified, it was opened. The first eligible patient in the morning would then be allocated to a laparoscopic appendectomy according to the passed-over envelope (D. Wall, written communication, June 9, 2000). If patients who present at night are sicker than those who present during the day, the resident's behavior would have biased the results against the open procedure.

This example demonstrates that, if those making the decisions about patient eligibility are aware of the treatment arm to which patients will be allocated (that is, if randomization is *unconcealed*), they may systematically enroll sicker, or less sick, patients in either the treatment or the control group. This behavior will defeat the purpose of randomization, and the study will yield a biased result^{16,17}. Careful investigators will ensure that randomization is concealed by having the medication prepared in a blinded fashion in a pharmacy; by employing remote randomization, in which the individual recruiting the patient makes a call to a methods center to discover the treatment arm to which the patient is allocated; or by making sure that the envelope containing the code remains sealed (which is, in our view, a much less secure approach).

Were all patients analyzed in the groups to which they were randomized?

Investigators can also ruin randomization by systematically excluding patients who do not receive the assigned treatment from the analysis of the results. Although it may seem that such patients should be excluded, this is not the case. The reasons that patients do not take their medication or do not receive a particular surgical intervention are often related to prognosis. In a number of randomized trials, patients who did not adhere to their treatment regimen fared worse than those who took their medication as instructed, even after all known prognostic factors had been taken into account and even when their medications were placebos¹⁸⁻²⁰. Excluding noncompliant patients from the analysis removes a group of patients with a worse prognosis, and the remaining patients will be destined to have a better outcome. Removing the non-compliers therefore destroys the unbiased comparison provided by randomization.

The situation is similar with regard to operative interventions. Some patients who are randomized to undergo surgery never have the operation because they are too sick or because they have an outcome that the operation was intended to prevent (such as stroke, deep venous thrombosis, or myocardial infarction) before they get to the operating room. If investigators include such patients, who are destined to have a poor outcome, in the control arm of a trial but not in the operative arm, even a useless operative therapy will appear to be effective. However, this apparent effectiveness will derive not from any benefit to those who had the operation but rather from the systematic exclusion of those with the poorest prognosis from the operative group. More commonly, however, patients randomized to the operative treatment arm do not receive the assigned treatment because of technical reasons. Again, these patients are likely to have poorer outcomes. As a result, investigators exclude these patients from the analysis, thereby losing the balance among prognostic factors that was achieved through randomization²¹⁻²³.

Because anything that happens after randomization can affect the chance that a patient will experience a specific event, it is important that all patients (even those who receive the wrong treatment) are analyzed in the groups to which they were initially randomized. This strategy, referred to as the *intention-to-treat* principle (Fig. 1), preserves the value of randomization: prognostic factors that are known and those that are not known will be, on average, distributed equally in the two groups, and the observed effect will be only that due to the assigned treatment. In reviewing a report of a randomized trial, one should look for evidence that the investigators analyzed all patients in the groups to which they were randomized.

Were patients in the treatment and control groups similar with respect to known prognostic factors?

The purpose of randomization is to create groups for which the prognosis, with respect to the target outcome, is similar. Sometimes, through bad luck, randomization will fail to achieve this goal. The smaller the sample size, the more likely

it is that the trial will suffer from prognostic imbalance. Consider a trial for the evaluation of a new osteoinductive agent for fracture-healing, in which patients with both closed and high-grade open fractures are enrolled. Patients with open fractures have a much worse prognosis than do those with closed fractures. The trial is small, with only eight patients. One would not be surprised if, by chance, all four closed fractures happened to be randomized to the new treatment and all four high-grade open fractures, to the control group. Such a result would seriously bias the study in favor of the new treatment. Were the trial to enroll 800 patients, the chances would be much smaller that randomization would place all 400 closed fractures in the treatment arm. The larger the sample size, the more likely it is that randomization will achieve its goal of prognostic balance.

Investigators can check how successful randomization has been by examining the distribution of all prognostic factors in the treatment and control groups. Clinicians should look for a display of the prognostic features of the patients in both groups at the

commencement of the study; these characteristics are referred to as *baseline* or *entry* prognostic features. Although it will never be known whether there is similarity between the groups with regard to the unknown prognostic factors, one can be reassured when the known prognostic factors are well balanced.

The question here is not whether there are significant differences between the treatment groups with regard to the known prognostic factors (in a randomized trial, one knows in advance that any such differences occurred by chance, making the frequently cited *p* values unhelpful) but rather what the magnitude of these differences is. If the differences are large, the validity of the study may be compromised. The stronger the relationship between the prognostic factors and the outcome, and the greater the differences in distribution between groups, the more the differences will weaken the strength of any inference about treatment impact (that is, the surgeon can place less confidence in the results of the study). In larger trials, randomization can occur in so-called *blocks* (that is, with or with-

out a known variable that affects the results, such as whether the fracture was open or closed), or it can be stratified according to variables such as age-group, dominant extremity, and so on. Both techniques help to ensure a balance of prognostic variables between groups.

If *blocked* or *stratified* randomization has not been used, all is not lost if the treatment groups are not similar at baseline. Statistical techniques permit adjustment of the study results for baseline differences. One should look for documentation of similarity for relevant baseline characteristics, and if substantial differences exist it should be noted whether or not the investigators conducted an analysis that adjusted for those differences. When both unadjusted and adjusted analyses lead to the same conclusion, one can be justifiably confident in the validity of the results.

Did experimental and control groups retain a similar prognosis after the study started?

Blinding

Since there is confusion about the terminology related to blinding (triple-blind, double-blind, and single-blind), it is useful to be explicit about who is blinded in the course of a trial²⁴.

Did investigators avoid effects of patient awareness of allocation: were patients blinded? The best way of avoiding the psychological impact of treatment (placebo effect) is to ensure that patients are unaware of whether they are receiving the experimental treatment. For instance, investigators conducting a trial to evaluate a new bone cement could blind patients by creating identical-looking incisions and packaging for the cement and the placebo.

Were aspects of care that affect prognosis similar in the two groups: were clinicians blinded? Differences in patient care other than the intervention under study can bias the results. In the example of the calcium-phosphate cement trial, if patients in the treatment group received more intensive postoperative care than did those in the control group, the results would yield an overestimation of the treatment effect.

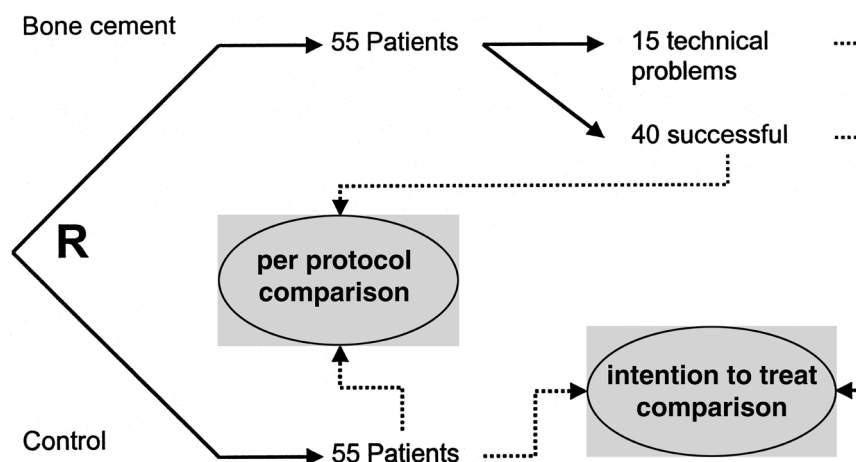


Fig. 1

Illustration of the intention-to-treat principle. Let us assume that fifteen patients who have been assigned to treatment of a radial fracture with supplementary use of bone cement have circumstances that make the injection of bone cement technically impossible. Excluding these patients and analyzing only those who actually received the treatment is called *per protocol analysis*. This, however, leads to an imbalance in baseline prognostic factors between groups, diminishing the effect of randomization. Intention-to-treat analysis takes into account the results for all patients who have been allocated to a particular treatment, thereby preserving the balance of prognostic factors from randomization. R = randomized.

Effective blinding eliminates the possibility of either conscious or unconscious differential administration of effective interventions to the treatment and control groups.

Was outcome assessed in a uniform way in experimental and control groups: were those assessing outcome blinded? If the treatment or the control group receives closer follow-up, target outcome events may be reported more frequently in that group. In addition, unblinded study personnel who are measuring or recording outcomes such as clinical status, quality of life, or radiographic findings may provide different interpretations of marginal results or offer differential encouragement during performance tests, either of which can distort the results²⁵. The study personnel who are assessing outcome can almost always be kept blinded, even if (as is the case for many operative therapies) the patient and the treating surgeon cannot. Investigators can take additional precautions by constructing a blinded adjudication committee to review clinical data and to decide issues such as whether a patient has a malunion, a nonunion, or another major complication. The more that judgment is involved in determining whether a patient has a target outcome, the more important blinding becomes; blinding is less crucial in studies in which the outcome is mortality due to any cause.

Was follow-up complete?

Ideally, investigators will know, at the conclusion of a trial, the status of each patient with respect to the target outcome. Patients whose status is unknown are often referred to as having been *lost to follow-up*. The greater the number of patients who are lost to follow-up, the more that a study's validity may be compromised. This is because patients who are lost to follow-up often have different prognoses from those who are not lost; the former group may be lost because they had an adverse outcome (including death) or because they were doing well and so did not return to the clinic to be assessed.

When does loss to follow-up seriously threaten validity? So-called rules of thumb (for example, a threshold of 20%) are misleading. Consider the hypothetical example of a ran-

domized trial in which 1000 patients are enrolled in both the treatment and the control group, with 200 patients (20%) (200 in the treatment group and 200 in the control group) subsequently being lost to follow-up. The treated patients have adverse outcomes at half the rate of the control patients (200 compared with 400), for a 50% reduction in relative risk. To what extent does the loss to follow-up potentially threaten our inference that treatment reduces the complication rate by half? If we assume the worst, that all treated patients lost to follow-up had the worst outcome, the number of adverse outcomes in the treatment group would be 400 (40%). If there were no adverse outcomes among the control patients who were lost to follow-up, our best estimate of the effect of treatment in reducing the rate of complications drops from $(1 - 200/400)$, or 50%, to $(1 - 400/400)$, or 0%. Thus, assuming the worst outcome does change the estimate of the magnitude of the treatment effect. If assuming a worst-case scenario does not change the inferences arising from the study results, then loss to follow-up is not a problem. If such an assumption significantly alters the results (as shown above), then validity is compromised.

Are the results of the study valid?

How well did the study of the calcium-phosphate cement¹ achieve the goal of creating groups with similar prognostic factors? The investigators stated that the study was randomized, but they did not explicitly address the issue of concealment. They documented the two groups' similarity with respect to age, initial radiographic displacement of the fracture, gender, hand dominance, and medications. They made no statement about blinding of patients, surgeons, or outcome assessors, nor did they make any explicit statement about loss to follow-up. However, all 110 patients appear to have been followed for twelve months. The second trial², in which treatment with a collagen-calcium phosphate material was compared with treatment with autogenous bone graft, was conducted with concealed randomization and blinding, but

TABLE II Validity Assessment of Calcium Phosphate-Cement Trials

	Sanchez-Sotelo et al. ¹ (Norian SRS)	Chapman et al. ² (Collagraft)
Date of report	2000	1997
Were patients randomized?	Yes	Yes
Was randomization concealed?	Not reported	Yes
Was an intention-to-treat analysis performed?	Not reported	Not reported
Were groups similar for baseline prognostic factors?	Yes	Yes
Were patients blinded?	No	Yes
Were surgeons blinded?	No	Yes
Were outcome assessors blinded?	No	Yes
Was follow-up complete?	100% at 1 yr	86% at 1 yr, 71% at 2 yr

there was a substantial loss of patients to follow-up (Table II).

The final assessment of validity is never a yes-or-no decision. Rather, one can think of validity as a continuum, ranging from strong studies that are very likely to yield an accurate estimate of the treatment effect to weak studies that are very likely to yield a biased estimate. Inevitably, the judgment as to where a study lies along this continuum involves some subjectivity. Since investigators will usually state that they have concealed randomization and blinded participants, it is likely that the validity of the calcium phosphate-cement trial was compromised by lack of concealment and blinding. In contrast, the collagen-calcium phosphate trial had limited bias because it utilized concealment of randomization and blinding, but 14% and 29% of the patients were lost to follow-up at one and two years, respectively.

Results

How large was the treatment effect?

Most investigators conducting randomized clinical trials carefully monitor how often patients experience adverse events or outcomes. Examples of these *dichotomous outcomes* (yes-or-no outcomes that either happen or do not happen) include reoperation, infection, and death. Patients either do or do not have an event, and the investigators report the proportion of patients who have such events.

Consider, for example, a study in which 20% (0.20) of the control group but only 10% (0.10) of the treatment group had an infection. How might these results be expressed? One way would be as the absolute difference (known as the *absolute risk reduction*, or *risk difference*) between the proportion who had an infection in the control group (X) and the proportion who had an infection in the treatment group (Y), or $X - Y = 0.20 - 0.10 = 0.10$. Another way to express the impact of treatment would be as a *relative risk*: that is, the risk of infection among patients receiving the new treatment compared with that among controls, or $Y/X = 0.10/0.20 = 0.50$.

The most commonly reported measure of dichotomous treatment effects is the complement of this relative risk, known as *relative risk reduction* (RRR). This measure is expressed as a percent: $(1 - Y/X) \times 100 = (1 - 0.50) \times 100 = 50\%$. A relative risk reduction of 50% means that the new treatment reduced the risk of infection by 50% compared with that among control patients; the greater the relative risk reduction, the more effective the therapy. Investigators may calculate the relative risk over a period of time, as in a survival analysis; this is called a *hazard ratio*.

How precise was the estimate of the treatment effect?

One can never know the true risk reduction; all that we have is the estimate provided by rigorous controlled trials, and the best estimate of the true treatment effect is that observed in the trial. This estimate is called a *point estimate* in order to remind us that, although the true value lies close to it, it is unlikely to be precisely correct. Investigators tell us the range within which the true effect likely lies by the statistical strategy of calculating *confidence intervals*.²⁶

Investigators usually (though arbitrarily) use the 95% confidence interval, which can be considered as defining the range that includes the true relative risk reduction 95% of the time. In other words, if the study were to be repeated 100 times, the point estimate of the result would be expected to lie within the confidence interval ninety-five of those 100 times. Investigators will seldom find the true relative risk reduction toward the extremes of this interval. Moreover, the true relative risk reduction will lie beyond these extremes only 5% of the time, a property of the confidence interval that is closely related to the conventional level of *statistical significance* of $p < 0.05$. The use of confidence intervals is illustrated in the following examples.

If a trial randomized 100 patients each to a treatment group and a control group, and if there were twenty malunions in the control group and fifteen in the treatment group, the authors would calculate a point estimate of 25% for relative risk reduction ($X = 20/100$ or 0.20, $Y = 15/100$ or 0.15, and $1 - Y/X = [1 - 0.75] \times 100 = 25\%$). One might guess, however, that the true relative risk reduction might be much smaller or much greater than 25% on the basis of a difference of just five malunions; in fact, one would be correct to surmise that the treatment might provide no benefit (a relative risk reduction of 0%) or that it might even cause harm (a negative relative risk reduction). Specifically, these results are consistent with both a relative risk reduction of -38% (that is, patients given the new treatment might be 38% more likely to have a malunion than control patients) and a relative risk reduction of nearly 59% (that is, patients given the new treatment might be almost 60% less likely to have a malunion than control patients). In other words, the 95% confidence interval for this relative risk reduction is -38% to 59%, and the trial has not really helped us to decide whether to offer the new treatment.

What if the trial enrolled not 100 but 1000 patients per group and the rates of malunion were the same as before; that is, there were 200 malunions in the control group ($X = 200/1000 = 0.20$) and 150 malunions in the treatment group ($Y = 150/1000 = 0.15$)? The point estimate of the relative risk reduction is 25% ($1 - Y/X = 1 - [0.15/0.20] \times 100 = 25\%$). In this larger trial, one might think that the true reduction in risk is much closer to 25% and, again, this would be correct. The 95% confidence interval for the relative risk reduction for this set of results is entirely on the positive side of zero and ranges from 9% to 41%.

These examples show that the larger the sample size of a trial, the larger the number of outcome events and the greater our confidence that the true relative risk reduction (or any other measure of efficacy) is close to what we have observed. In the second example above, the lowest plausible value for the relative risk reduction was 9% and the highest value was 41%. The point estimate—in this case, 25%—is the one value most likely to represent the true relative risk reduction. As one considers values farther and farther from the point estimate, these values become less and less consistent with the observed relative risk reduction. By the time that one crosses

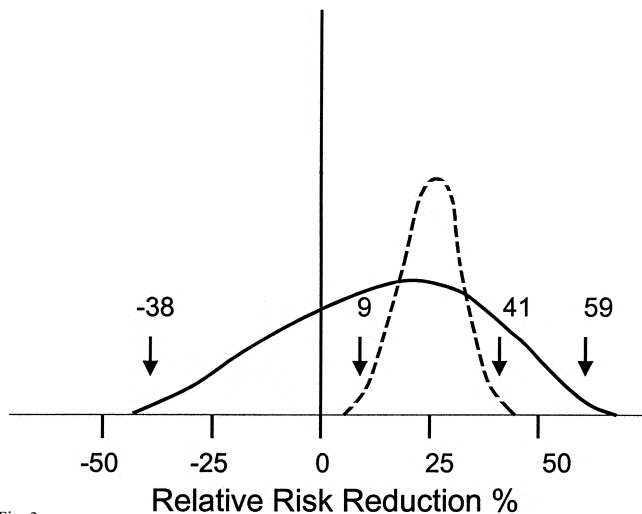


Fig. 2

Graph of the data from two studies with the same point estimate, a 25% relative risk reduction, but different sample sizes and correspondingly different confidence intervals. Larger confidence intervals are associated with smaller trials (solid line).

the upper or lower boundary of the 95% confidence interval (9% to 41%), the values are extremely unlikely to represent the true relative risk reduction, given the point estimate (that is, the observed relative risk reduction).

Figure 2 represents the confidence intervals around the point estimate of a relative risk reduction of 25% in these two examples, with a risk reduction of 0 representing no treatment effect. In both scenarios, the point estimate of the relative risk reduction is 25%, but the confidence interval is far narrower in the second scenario (because of a much larger sample size).

It is evident that the larger the sample size, the narrower the confidence interval. How can a clinician ascertain if a study is large enough to allow confidence in a conclusion? In a *positive* study—a study in which the authors conclude that the treatment is effective—one can look at the lower boundary of the confidence interval. In the second example, this lower boundary was 9%. If this relative risk reduction (the lowest relative risk reduction that is consistent with the study results) is still clinically important (that is, if it is large enough for the surgeon to recommend the treatment to the patient), then the investigators have enrolled a sufficient number of patients. If, on the other hand, a relative risk reduction of 9% is not considered important, then the study cannot be considered definitive, even if the results are statistically significant (that is, if they exclude a risk reduction of 0). Keep in mind that the probability of the true value being less than the lower boundary of the confidence interval is only 2.5% and that a different criterion for the confidence interval (a 90% confidence interval, for instance) might be as (or more) appropriate.

The confidence interval also helps us to interpret a *negative* study—one in which the authors have concluded that the experimental treatment is no better than the control therapy²⁷. All one needs to do is to examine the upper boundary of the

confidence interval. If the relative risk reduction at this upper boundary would, if true, be clinically important, then the study has failed to exclude an important treatment effect. For example, consider the first scenario presented in this section, the study with 100 patients in each group. This study does not exclude the possibility of harm (indeed, it is consistent with a 38% increase in relative risk), the associated *p* value would be greater than 0.05, and the study would be considered *negative* in that it failed to show a convincing treatment effect (Fig. 2). Recall, however, that the upper boundary of the confidence interval was a relative risk reduction of 59%. The study has clearly failed to exclude an important beneficial treatment effect.

What can the clinician do if the confidence interval for the relative risk reduction is not reported? There are three possible approaches. The easiest approach is to examine the *p* value. If the *p* value is exactly 0.05, then the lower boundary of the 95% confidence interval for the relative risk reduction has to lie exactly at zero (a relative risk of 1), and one cannot exclude the possibility that the treatment has no effect. As the *p* value decreases below 0.05, the lower boundary of the 95% confidence interval for the relative risk reduction rises above zero.

The second approach, involving some quick arithmetic, can be used when the study includes the value for the standard error of the relative risk reduction (or of the relative risk). This is because the upper and lower boundaries of the 95% confidence interval for a relative risk reduction are the point estimate plus and minus twice this standard error (relative risk reduction $\pm 2 \times$ standard error).

The third approach involves calculating the confidence intervals oneself²⁸ or asking someone else (such as a statistician) to do so. Once the confidence intervals are obtained, it is known how high and low the relative risk reduction might be; that is, the precision of the estimate of the treatment effect is known and it is possible to interpret the results as described above.

Not all randomized trials have dichotomous outcomes, nor should they. For example, the authors of the Norian SRS study¹ reported differences in pain according to a visual analog scale as well as in grip strength in both the treatment (Norian SRS) group and the control group. Both pain and grip strength are continuous variables. The mean grip strength (expressed as the percentage of that on the normal side) at one year was 92% in the Norian SRS group compared with 80% in the control group. The mean difference in grip strength was 12% in favor of the patients treated with Norian SRS.

Here, too, one should look for the 95% confidence interval for this difference in grip strength and consider the implications. The lower boundary of the 95% confidence interval is 9% and the upper boundary is 15%. Thus, even the lower boundary of the confidence interval favors the treatment group and the difference is still clinically important.

Having determined the magnitude and precision of the treatment effect, clinicians can turn to the final question of how to apply the results of the study to their patients and their clinical practice.

TABLE III Results of the Two Trials in Terms of Malunion

Trial	Treatment Group		Control Group		Risk Reduction (95% Confidence Interval) (%)	Risk Difference* (%)	No. of Patients Needed to Treat†
	No. of Events/ No. of Patients or Fractures	Rate of Malunion (X) (%)	No. of Events/ No. of Patients or Fractures	Rate of Malunion (Y) (%)			
Sanchez-Sotelo et al. ¹	10/55	18	23/55	42	57 (17-77)	24	4
Chapman et al. ²	4/117	3.4	10/132	7.6	55 (-26-80)	4.2	24

*The risk difference (RD) is calculated as $Y - X$. †The number of patients needed to treat (that is, the number of patients who must be treated with bone cement in order to prevent one malunion) is calculated as $1/\text{RD}$.

Results of the calcium phosphate-cement trial

A malunion developed in ten (18%) of the fifty-five patients in the treatment group compared with twenty-three (42%) in the control group. (Malunion was reported as a dichotomous variable.) This meant that the relative risk of malunion with treatment was 0.43 (18/42) and the relative risk reduction was 57% ($1 - 0.43$). The 95% confidence interval for the relative risk reduction is 17% to 77% (Table III). Complications of cement use included a 70% rate of extrusion of cement beyond the fracture site (not clinically important) and a 1% rate of reoperation due to intra-articular extrusion of bone cement (clinically important).

In the collagen-calcium phosphate trial², 249 fractures of long bones (the femur, humerus, radius, ulna, and tibia) were followed for more than two years. Patients were randomized to internal or external fixation supplemented either with Collagraft and autogenous bone marrow (obtained from fine-needle aspiration) or with autogenous iliac-crest bone graft. The rates of malunion (deformity) associated with Collagraft and bone graft were 3.4% and 7.6%, respectively. This represents a relative risk reduction of 55% for malunion in association with Collagraft, although the 95% confidence interval is wide (-26% to 80%). At one extreme the use of Collagraft reduced the risk of malunion by 80%, and at the other extreme it actually increased the risk. Autogenous graft was associated with an overall infection rate of 14.2%, whereas Collagraft was associated with an infection rate of 4.9%.

Applicability/Generalizability

Can the results be applied to my patient?

Often, the patient whom you must treat is somewhat different from those enrolled in a reported trial. If the patient would have been eligible for the study—that is, if the patient meets all of the inclusion criteria and none of the exclusion criteria—then you can apply the results to your patient's care with considerable confidence.

Even here, however, there is a limitation: treatments are not uniformly effective. Typically, some patients respond extremely well, while others derive no benefit. Conventional randomized trials estimate mean treatment effects; thus, the clinician will likely be exposing some patients to the cost and risks of the treatment without benefit. Additionally, whenever

there is clinical skill involved in carrying out the treatment under consideration, the surgeon must ask if his or her individual level of skill with the treatment is likely to be comparable with that of the surgeons who provided the care in the reported trial.

A final issue arises when your patient shares the features of a subgroup of patients in the reported trial. In assessing the results of a trial (especially when the treatment does not appear to have been efficacious for the average patient), the investigators may have examined a large number of subgroups of patients with different stages of an illness, different comorbid conditions, and different ages at the time of entry into the trial. Quite often these subgroup analyses were not planned ahead of time, and the data are simply dredged in an attempt to find an effect. Investigators may sometimes overinterpret these *data-dependent* analyses as demonstrating that the treatment really has a different effect in a subgroup of patients; for instance, it may be suggested that patients who were older or sicker benefited substantially more or less than did other subgroups of patients in the trial.

One should be skeptical of subgroup analyses^{29,30}. The treatment is likely to benefit the subgroup more or less than the other patients only if the difference in the effects of treatment among subgroups is large and is very unlikely to have occurred by chance. Even when these conditions apply, the results may be misleading if the investigators did not specify their hypotheses before the study began, if they had a very large number of hypotheses, or if other studies failed to replicate the findings.

Were all clinically important outcomes considered?

Treatments are indicated when they provide important benefits. The demonstration that a new orthopaedic implant increases the range of motion of a joint does not necessarily mean that this implant should be adopted for routine use, particularly if there is no evidence that an increased range of motion results in important functional improvement. What is required is evidence that the treatment improves outcomes that are important to patients, such as reducing the rate of reoperation due to infection, malunion, or nonunion; improving function; or increasing the rate of survival.

Another long-neglected outcome is that of the resource

implications of alternative treatment strategies. Few randomized trials measure either direct costs, such as drug or program expenses and health-care-worker salaries, or indirect costs, such as the patient's loss of income due to illness or complications. Nevertheless, the increasing constraints on resources that health-care systems face mandate careful economic analysis, particularly of resource-intensive interventions.

Are the likely treatment benefits worth the potential harm and costs?

If one can apply the study's results to his or her patient, and if its outcomes are clinically important, the next question concerns whether the probable treatment benefits are worth the effort that the surgeon and the patient must put into the enterprise. A 25% reduction in the relative risk of infection may sound quite impressive, but its impact on the patient and the surgeon's practice may nevertheless be minimal. This concept is illustrated with use of a concept known as *number needed to treat (NNT)*³¹, which is the inverse of the absolute risk difference (1/risk difference).

Consider the following illustration of the number-needed-to-treat concept, based on data from a recent systematic review of randomized trials comparing the use of reamed and nonreamed intramedullary nailing in 350 patients who had long-bone fractures of the lower extremity³². The authors reported that 5% of the patients treated with reamed nailing and 15% of those treated with nonreamed nailing had a nonunion. This translates into a relative risk of nonunion of 0.33 (95% confidence interval, 0.16 to 0.68) and a relative risk reduction of 67% (95% confidence interval, 32% to 84%) in association with reamed intramedullary nailing. The risk difference of 10% (15% – 5%) suggests that, for every ten patients treated with reamed intramedullary nailing, the surgeon can prevent one nonunion (number needed to treat = $1/0.10 = 10$.)

While reamed intramedullary nailing seems to be an attractive alternative to nonreamed intramedullary nailing when nonunion rates are considered, the obvious drawback is the potential for increased infection with reamed canal preparation in patients with open fractures. The rate of infection rarely exceeds 3% in patients with closed tibial fractures and is at least four times higher (12%) in patients with open fractures who are treated with reamed nailing^{21,22}. Thus, one might expect that for every 100 patients whom the surgeon might consider treating with reamed nailing, ten nonunions would be prevented at the cost of nine infections (risk difference = 12% – 3% = 9%; number needed to treat = 11). The utility of reamed nailing for the treatment of open fractures suddenly becomes less certain.

Resolution of the Scenario

The calcium phosphate-cement trial leaves us with several important questions concerning its methodology. It is unlikely that randomization was concealed, and without concealment the surgeon may have been able to determine the treatment arm to which the patient would be allocated.

Clearly, knowledge of patient allocation can result in the exclusion of patients who are deemed not likely to benefit from the therapy. For instance, if a surgeon believes that bone cement is needed for the treatment of severely displaced fractures, he or she may be inclined to find a reason to exclude a patient whom he or she knows will be randomized to the placebo group.

The authors do not tell us if all patients were analyzed in the groups to which they were originally randomized (intention to treat); however, this is probably not important as there were evidently no crossovers in treatment (that is, all patients received the treatment to which they were randomized). Lack of blinding of outcome assessors further limits the study's validity. The methodological strengths of the calcium phosphate-cement trial lie in its randomized design and its completeness of follow-up (100%). The apparent omission of independent assessment of the radiographic outcome (malunion) adds a serious potential bias to the results in favor of the bone cement, given that the authors endorse the product being studied.

Table III indicates that, if the results were valid (which we doubt), surgeons must use calcium-phosphate bone cement in four patients to prevent a malunion of the distal aspect of the radius in one of them. However, the authors also report a 70% rate of cement extrusion into the soft tissues (with 1% intra-articular extrusion requiring reoperation). Therefore, for every 100 patients whom one might consider treating with this new bone cement, it should be possible to prevent twenty-five malunions at the cost of seventy patients having soft-tissue extrusion and one patient requiring a reoperation. Clearly, the choice is not a simple one. One must balance the clinical impact of malunion on the patient's function and quality of life with the impact of extruded cement. If one assumes that functional outcome is highly correlated with radiographic malunion, then the use of this bone cement may, in fact, be justified. As it turns out, these same investigators report a significant association between radiographic parameters and functional outcomes. Moreover, they state that half of the extruded cement disappears within a few years and that it causes only transient discomfort in most patients (89%). Given this information, if the results represent an unbiased estimate, the apparent benefits of Norian SRS outweigh its disadvantages. The likelihood of bias, however, leaves us in considerable doubt.

Can the results of this trial be generalized to any fifty-five-year-old female patient with a distal radial fracture? On the basis of the eligibility criteria put forth by the authors, you determine that your hypothetical patient would have been eligible for inclusion in the calcium phosphate-cement trial. Bearing in mind the limitations in validity, you can therefore be reasonably confident in applying these results to your patient's care.

The strengths of the study on collagen calcium-phosphate cement include concealment of allocation and blinding of outcome assessors. Its weaknesses include the omission of an intention-to-treat analysis. The study's main finding is a 55%

reduction in the risk of malunion (95% confidence interval, –26% to 80%) with use of Collagraft in the treatment of long-bone fractures. The authors report a *negative* trial, with no difference in malunion rates between groups. However, the upper limit of the 95% confidence interval (if true) would be highly persuasive evidence in favor of Collagraft, particularly given the absence of donor-site morbidity (infection and pain in association with the iliac-crest grafts) noted with use of this material. Clearly, the sample size is too small to allow the claim that there is no difference between the two treatments. In fact, the point estimate suggests that collagen calcium-phosphate cement may be superior to autogenous bone graft in maintaining fracture alignment. Larger trials with more patients are needed to resolve this issue.

Despite its potentially promising results, this study² may not be entirely generalizable to your patient. Of the 249 fractures, only 18% involved the distal aspect of the radius. Moreover, eligibility for enrollment in this study would have necessitated the decision to use a bone graft and either internal or external fixation. Ultimately, this study is completely inapplicable to your patient.

In conclusion, once the surgeon has found an article of interest on an orthopaedic surgical intervention, it is necessary to assess the quality of the evidence therein. To the extent that the quality is poor, the inferences that are drawn from the study will be weakened; however, if the quality is acceptable, one must determine the range (95% confidence interval) within which the true treatment effect lies. Then, one must consider if the result can be generalized to one's own patient and whether the investigators have provided information about all clinically important outcomes. Finally, it is neces-

sary to compare the relative benefits and risks of the intervention. If the benefits appear to outweigh the risks, then the intervention may be useful for one's patient.

Given the time constraints of busy surgical practices and surgical training programs, applying this analysis to every relevant article will be challenging. However, the basic steps of this process are essentially what we all do hundreds of times each week when treating patients. Making this process explicit, with guidelines to assess the strength of the available evidence, will serve to improve patient care. It also will allow us to defend therapeutic interventions on the basis of available evidence rather than anecdotal information. ■

NOTE: Concepts in this article have been taken, in part, from the *User's Guide to the Medical Literature*, edited by Gordon H. Guyatt and Drummond Rennie.

M. Bhandari, MD, MSc
G.H. Guyatt, MD, MSc
Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Center, Room 2C12, 1200 Main Street West, Hamilton, ON L8N 3Z5, Canada. Please address requests for reprints to G.H. Guyatt. E-mail address for M. Bhandari: bhandari@sympatico.ca

M.F. Swiontkowski, MD
Department of Orthopaedic Surgery, University of Minnesota, Box 492, Delaware Street N.E., Minneapolis, MN 55455

The authors did not receive grants or outside funding in support of their research or preparation of this manuscript. They did not receive payments or other benefits or a commitment or agreement to provide such benefits from a commercial entity. No commercial entity paid or directed, or agreed to pay or direct, any benefits to any research fund, foundation, educational institution, or other charitable or nonprofit organization with which the authors are affiliated or associated.

References

1. Sanchez-Sotelo J, Munuera L, Madero R. Treatment of fractures of the distal radius with a remodelable bone cement: a prospective, randomised study using Norian SRS. *J Bone Joint Surg Br.* 2000;82:856-63.
2. Chapman MW, Bucholz R, Cornell C. Treatment of acute fractures with a collagen-calcium phosphate graft material. A randomized clinical trial. *J Bone Joint Surg Am.* 1997;79:495-502.
3. Haynes RB, Mukherjee J, Sackett DL, Taylor DW, Barnett HJ, Peerless SJ. Functional status changes following medical or surgical treatment for cerebral ischemia. Results of the extracranial-intracranial bypass study. *JAMA.* 1987;257:2043-6.
4. Crette S, Marcoux S, Truchon R, Grondin C, Gagnon J, Allard Y, Latulippe M. A controlled trial of corticosteroid injections into facet joints for chronic low back pain. *N Engl J Med.* 1991;325:1002-7.
5. Xamoterol in severe heart failure. The Xamoterol in Severe Heart Failure Study Group. *Lancet.* 1990;336:1-6.
6. Packer M, Carver JR, Rodeheffer RJ, Ivanhoe RJ, DiBianco R, Zeldis SM, Hendrix GH, Bommer WJ, Elkayam U, Kukin ML, et al. Effects of oral milrinone on mortality in severe chronic heart failure. The PROMISE Study Research Group. *N Engl J Med.* 1991;325:1468-75.
7. Flosequin withdrawn. *Lancet.* 1993;342:235.
8. Hampton JR, van Veldhuisen DJ, Kleber FX, Cowley AJ, Ardia A, Block P, Cortina A, Cserhalmi L, Follath F, Jensen G, Kayanakis J, Lie KI, Mancía G, Skene AM. Randomised study of effect of ibopamine on survival in patients with advanced severe heart failure. Second Prospective Randomised Study of Ibopamine on Mortality and Efficacy (PRIME II) Investigators. *Lancet.* 1997;349:971-7.
9. Califf RM, Adams KF, McKenna WJ, Gheorghiade M, Uretsky BF, McNulty SE, Darius H, Schulman K, Zannad F, Handberg-Thurmond E, Harrell FE Jr, Wheeler W, Soler-Soler J, Swedberg K. A randomized controlled trial of epoprostenol therapy for severe congestive heart failure: The Flolan International Randomized Survival Trial (FIRST). *Am Heart J.* 1997;134:44-54.
10. Sacks HS, Chalmers TC, Smith H Jr. Sensitivity and specificity of clinical trials. Randomized v historical controls. *Arch Intern Med.* 1983;143:753-5.
11. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med.* 1983;309:1358-61.
12. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. *Stat Med.* 1989;8:441-54.
13. Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Control Clin Trials.* 1990;11:339-52.
14. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ.* 1998;317:1185-90.
15. Hansen JB, Smithers BM, Schache D, Wall DR, Miller BJ, Menzies BL. Laparoscopic versus open appendectomy: prospective randomized trial. *World J Surg.* 1996;20:17-20; discussion 21.
16. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA.* 1995;273:408-12.
17. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet.* 1998;352:609-13.
18. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. *N Engl J Med.* 1980;303:1038-41.
19. Asher WL, Harper HW. Effect of human chorionic gonadotrophin on weight

- loss, hunger, and feeling of well-being. *Am J Clin Nutr*. 1973;26:211-8.
20. **Hogarty GE, Goldberg SC.** Drug and psychotherapy in the aftercare of schizophrenic patients. One-year relapse rates. *Arch Gen Psychiatry*. 1973;28:54-64.
 21. **Keating JF, O'Brien PJ, Blachut PA, Meek RN, Broekhuysen HM.** Locking intramedullary nailing with and without reaming for open fractures of the tibial shaft. A prospective, randomized trial. *J Bone Joint Surg Am*. 1997;79:334-41.
 22. **Blachut PA, O'Brien PJ, Meek RN, Broekhuysen HM.** Interlocking intramedullary nailing with and without reaming for the treatment of closed fractures of the tibial shaft. A prospective, randomized study. *J Bone Joint Surg Am*. 1997;79:640-6.
 23. **Chapman JR, Henley MB, Agel J, Benca PJ.** Randomized prospective study of humeral shaft fracture fixation: intramedullary nails versus plates. *J Orthop Trauma*. 2000;14:162-6.
 24. **Devereaux PJ, Manns BJ, Ghali WA, Quan H, Lacchetti C, Montori VM, Bhandari M, Guyatt GH.** Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. *JAMA*. 2001;285:2000-3.
 25. **Guyatt GH, Pugsley SO, Sullivan MJ, Thompson PJ, Berman L, Jones NL, Fallen EL, Taylor DW.** Effect of encouragement on walking test performance. *Thorax*. 1984;39:818-22.
 26. **Altman DG, Gore SM, Gardner MJ, Pocock SJ.** Statistical guidelines for contributors to medical journals. In: Gardner MJ, Altman DG, editors. *Statistics with confidence. Confidence intervals and statistical guidelines*. London: British Medical Journal; 1989. p 83-100.
 27. **Detsky AS, Sackett DL.** When was a "negative" trial big enough? How many patients you needed depends on what you found. *Arch Intern Med*. 1985;145:709-12.
 28. **Sackett DL, Haynes RB, Guyatt GH, Tugwell P.** *Clinical epidemiology: a basic science for clinical medicine*. 2nd ed. Boston: Little, Brown; 1991. p 218.
 29. **Oxman AD, Guyatt GH.** A consumer's guide to subgroup analyses. *Ann Intern Med*. 1992;116:78-84.
 30. **Assmann SF, Pocock SJ, Enos LE, Kasten LE.** Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000;355:1064-9.
 31. **Laupacis A, Sackett DL, Roberts RS.** An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med*. 1988;318:1728-33.
 32. **Bhandari M, Guyatt GH, Tong D, Adili A, Shaughnessy SG.** Reamed versus nonreamed intramedullary nailing of lower extremity long bone fractures: a systematic overview and meta-analysis. *J Orthop Trauma*. 2000;14:2-9.